

**Project White Paper**

Grant number: HG-50050-13

Title of Project: Digital Corpus of Literary Papyri

Name of Project Director: Roger S. Bagnall

Name of Grantee Institution: New York University

Date Submitted: 1 December 2017

## Introduction: Background and Intent

The **Digital Corpus of Literary Papyri** initiative aims to develop a comprehensive, searchable digital collection of literary and subliterate texts surviving on papyrus and similar materials from the area of ancient Egypt and environs. Papyrologists have long made the distinction between literary and documentary texts in their discipline, with the latter comprising day-to-day writing like contracts, tax returns, administrative or personal correspondence, and the like. Literary papyri are further often subdivided in order to distinguish between "high literature" and "sub" or "paraliterary" treatises related to science, education, religion, and popular belief. The grant proposal estimated the total number of surviving literary and subliterate texts in Greek, Latin, and various forms of Egyptian at approximately 13,000, but our subsequent work has brought the number up over 14,500 (see below). The DCLP collection of literary and subliterate papyri is intended to complement the existing digital corpus of documentary texts published at <http://papyri.info>, which at time of writing number approximately 55,000.

The work reported here was funded by two linked grants: one from the National Endowment for the Humanities to New York University (NEH HG-50050-13) and the other from the German Research Foundation (*Deutsche Forschungsgemeinschaft*) to the Institute of Papyrology at the University of Heidelberg (DFG GZ: AS 382/2-1). The proposals for these grants were developed with support from a joint DFG-NEH planning grant (NEH HW-50021-11, executed between September 2011 and August 2012).

## Project Activities

### Changes in Activities and Personnel

The project has been successful in delivering on major goals (see "Accomplishments," below), despite having needed to restructure radically the work plan and schedule outlined in the original proposal. The NYU team had expected to be able to establish a complete clone of the papyri.info system during the first few months of the project and to update it frequently as code changes were made so that project participants and third parties could assess both the pace and the direction of the work. Because of personnel issues and technical challenges, this initial work package took almost two years to complete, rather than the anticipated 4+/- months. In the meantime, the Heidelberg team moved ahead on the parsing and refactoring of metadata from the Leuven Database of Ancient Books while the NYU team put in place and used a separate development and test environment for key display components.

As previously reported in our semi-annual performance updates, the initial work plan assumed the availability and full-time commitment of an individual employed in the NYU libraries who was deeply familiar with the code and operation of papyri.info, which was then jointly hosted by the libraries and ISAW. Near the time of project award, this individual took new employment with the Duke Collaboratory for Classics Computing (DC3), which had just been founded, and papyri.info hosting and development was transferred there. Duties of the new position precluded anything but a minor consulting role on DCLP for that individual (although this consulting has been enormously valuable), and accordingly we submitted and were granted budget and schedule modifications necessary to identify and hire a replacement. Unfortunately, the combination of

unusual technical skill requirements, experience level, visa restrictions, and limited term made it impossible to identify a suitable candidate. The gap was ultimately filled with a combination of graduate student labor, independent contractors, a restructuring of the Project Manager's technical responsibilities, and consulting from DC3.

The second factor affecting the project timeline and ordering of work was the unexpected difficulty of setting up the development clone of the papyri.info code. The system requires a large number of third-party, open source components and includes dependencies on specific and sometimes no-longer-supported versions of same. These dependencies and associated configuration steps were not always fully documented. Without the full-time attention of an individual already well-versed in the entire software stack and its deployment, much trial and error was required. When initial efforts to establish the development server on a U.S.-based leased Linux host finally reached an advanced stage, it became clear that the Heidelberg team would have difficulty working with it effectively because of the slow pace of data transfer between their workstations and the server. Accordingly, the effort began again on a development server in Heidelberg, but progress there was ultimately thwarted by the fact that the Heidelberg server was running the OSX operating system, instead of Linux, and key components and services needed were not available therein. Ultimately, Heidelberg obtained access to a Linux development server and the two teams collaborated to get it up and running. It is this system that currently provides the demonstration version of the DCLP at <http://litpap.info>. DC3 and papyri.info leadership were kept aware of problems as they developed and in many cases provided troubleshooting help, code updates, and direct intervention in order to resolve them. These experiences are informing decision-making about the next-generation software architecture for papyri.info and other resources, which is now being designed.

The Data Management Plan for this project called for the preservation of data in a digital preservation repository at NYU alongside the documentary papyri. In writing this plan, we had anticipated that NYU libraries would have completed steps to accession the papyri.info data into their institutional repository and therefore that adding the DCLP data would be straightforward. The transfer of papyri.info and its lead developer to DC3, combined with administrative and technical changes in library systems and services, put an end to this plan. ISAW's Digital Programs and Libraries team has since been developing -- external to the effort on this grant-funded project -- tools and specifications for depositing research data and imagery collections into the NYU Faculty Digital Archive; it is there that we expect to place the DCLP data in due time (see below, sub "Continuation of the Project"). Meantime, all code and data is replicated on multiple machines in New York, Heidelberg, and Huntsville, Alabama, and on the commercial Github.com website.

## **Software Development**

Despite the staffing and schedule changes described above, all software development activities addressed in the original proposal's work packages were carried out within budget. Software modifications were made to all major components of the papyri.info code base, including the Papyrological Navigator (PN; which provides for search and display of metadata and editions)

and the Papyrological Editor (PE; which facilitates the actual creation and editing of metadata and editions, as well as a review process workflow).

Early in the period of performance, the NYU Project Manager established an online issue tracker in order to facilitate requirements definition and implementation tracking. This tool proved essential to organizing and prioritizing technical work across the teams; however, on-going unplanned effort was required to capture relevant information from email threads, Skype conversations, and in-person meetings into the tracker; to keep each ticket thread focused on a single issue; and to secure implementation decisions from project stakeholders and subject-matter experts. Available resources proved inadequate to manage these social processes on a day-by-day basis; consequently, sometimes essential requirements were articulated (indeed, some were discussed extensively) without being effectively described, prioritized, and surfaced to developers in a timely manner. Conversely, some problems and delays encountered by developers were not surfaced to colleagues and consultants as early or effectively as they might have been. Although some of these difficulties can certainly be attributed to the distances (geography, time zones, institutional calendars, native languages) between the various team members, two lessons for future efforts may be gleaned from our experience: first, project plans and budgets should include an experienced individual to manage the tracker, train team members on its use, and capture relevant content from parallel interaction streams; and second, that project leadership should articulate, implement, and enforce a collaborative social contract that explains the processes and artifacts the project uses for requirements definition and tracking.

Both project partners contributed to the actual software work, with the Heidelberg team taking the lead on modifications to the PE and its underlying workflow module, and to the browse mechanisms in the PN. The NYU team focused on extensions to the code used to transform and display metadata and editions in the PN, and to the computational grammar used in the PE to convert user-friendly shorthand codes for editorial observations and supplements to texts into the standard encoding used natively in the dataset. The majority of this NYU effort focused on the addition of features necessary for literary papyrology in response to requirements defined by collaborators and prioritized by members of the project team.

Under subaward from NYU, staff at the DC3 provided technical support to both teams; fully rewrote and thoroughly integrated the new search and browse functionality for scalability and effectiveness; and conducted a full code review and consolidation. The primary object of this latter activity constitutes another key deliverable identified in the original proposal: re-integration of DCLP code modifications into the core papyri.info codebase so that literary and subliterate materials can be published, searched, and curated alongside the documentary.

## **Encoding of Texts and Metadata**

An essential element of this project has been the encoding of descriptive data (metadata) and text (editions) using the EpiDoc customization of the Text Encoding Initiative XML tagset. The Heidelberg team carried out the conversion of metadata from the Leuven Database of Ancient Books and other sources, supported by the NYU team as needed for research, testing, and

validation of encoding structures. Edition encoding was carried out by various collaborators, including participants in training and evaluation sessions (see below). See the "Accomplishments" section for further information.

## **Training and Evaluation Sessions**

During the course of the project, DCLP team members led a number of training sessions aimed at instructing scholars in the use of the projects editorial system and user interface.

### **Parma, Italy (September 2014)**

This event was hosted by Professor Andorlini in the context of the ERC-funded DigMedText project mentioned below, sub "Accomplishments". At the time there was no free-standing DCLP content creation tool, so participants learned to enter the texts in the papyri.info system; the XML was subsequently transferred to Github and from there to DCLP.

### **Würzburg, Germany (March 2017)**

This was the first training session to use the modified DCLP editorial system and was funded from the DFG grant. The event involved four instructors, nine students, and one programmer. Participants learned how to curate and supplement the metadata and enter transcriptions of Greek texts using Leiden+. Results were very good and the session generated interest in a follow-up workshop concentrating on more advanced concepts such as editorial roles and peer review.

### **Heidelberg, Germany (June 2017)**

This advanced workshop extended the training and evaluation conducted in Würzburg and was also funded from the DFG grant.

### **New York (August 2017)**

This workshop, funded out of the NEH grant, brought together 6 participants (including the NYU PI and Project Manager) with 2 instructors (Valeria Piano from the University of Florence and Nicola Reggiani from the University of Parma) for 2 days of intensive work with the modified PE and PN. A number of challenging textual aspects were explored and successfully tested, and important refinements to metadata editing and workflow management were identified for future implementation.

## **Outreach**

In addition to the sessions reported above, Ast (Heidelberg PI) and Essler (Würzburg) held multiple lectures on the topic of the DCLP project (see below). Moreover, two semester-long online classes were offered to students around the world in the summer semester 2016 and 2017. The first, entitled "Online Approaches to Editing Greek Literature. The Philosophical Papyri," was taught by Ast and Essler and was devoted to literary papyrology, with a focus on philosophical texts. Among other things, the instructors introduced students to DCLP and assigned projects that involved the entry of new transcriptions (for an example, see the fragment of a philosophical treatise at <http://litpap.info/dclp/63976>). Lougovaya (Heidelberg collaborator and DCLP editor) also taught a webinar on literary texts in the summer semester 2017. It was entitled "Greek Literary Ostraca" and resulted in student contributions of nine new transcriptions of Greek literature preserved on potsherds, including an erotic epigram from the

island of Rhodes (<http://litpap.info/dclp/217929>) and an anthology of Greek literature (<http://www.litpap.info/dclp/62676>).

## Lectures

- Ast, R., "Digital Corpus of Literary Papyri," University of Toronto. Conference Panel: "Ancient MakerSpaces," Society for Classical Studies, Toronto, January 7, 2017
- Ast, R., "A Digital Corpus of Literary Papyri," Parma University. Conference: "Greek Medical Papyri - Text, Context, Hypertext," November 2-4, 2016
- Ast, R., "A Digital Corpus of Literary Papyri," Panel: Digital Humanities in Biblical, Early Jewish, and Christian Studies; Papyrology and Early Christian Backgrounds, Society for Biblical Literature, Atlanta, GA, November 23, 2015
- Ast, R., "Editions of Ancient Papyri Born Digital. Aspirations and Current Obstacles," Classics Colloquium, University of Pennsylvania, September 24, 2015
- Ast, R. - H. Essler, "Digital Corpus of Literary Papyri (DCLP)," Der 8. Deutsche Papyrologentag, Berlin, July 18-20, 2014
- Essler, H., "Anagnosis und DCLP," Parma University. Conference: "Greek Medical Papyri - Text, Context, Hypertext," November 2-4, 2016
- Essler, H., "Anagnosis und DCLP," Der 9. Deutsche Papyrologentag, Leipzig, June 23-25, 2017

## Accomplishments

The proposal narrative for the NEH grant described the project scope as follows:

The Digital Corpus of Literary Papyri (DCLP) will be aimed at two goals. First, it will **build the technology framework** for the texts, metadata and translations of literary papyri, relying on core technology that is already in place for documentary papyri at [www.papyri.info](http://www.papyri.info); secondly, it will **stress-test this framework** by incorporating into the DCLP the results of several existing projects as well as facilitating the work of a handful of new ones. By the end of the funded period, we intend to have a fully functional, searchable and editable set of texts in place delivered by a system that has been customized to some of the key requirements of literary papyrology (e.g., extended apparatus capability and accommodation of sigla found predominantly in literary texts). This preliminary corpus of texts will be part of a scalable infrastructure that will allow ongoing, largely volunteer-driven data curation and supplementation, on the model of existing documentary projects.

The joint project team has succeeded in producing the promised "technology framework" (a modified version of the papyri.info code) and in stress-testing it with a range of real-world metadata and editions.

The web application is currently up and running at <http://litpap.info/>, demonstrating functional browsing and curated editing of descriptive information (metadata) about the 14,620 literary and subliterate papyri so far cataloged, and of nearly 920 scholarly transcriptions (editions) of the texts themselves at the time of writing. In order to realize the technical framework of litpap.info,

software modifications have been made to every major component of the codebase inherited from papyri.info, as outlined in the "Project Activities" section, above.

The original proposal also articulated the goal of re-integrating DCLP code modifications into the core papyri.info codebase so that literary and subliterary materials can be published, searched, and curated alongside the documentary in a single system (i.e., at papyri.info). It has not been possible to achieve this goal during the reporting period, due to the the need to finish developing support for translation editing and batch creation of metadata; however, work on these elements continues at Heidelberg under separate funding and DC3 anticipates successful integration and deployment of same once complete (see below, sub "Continuation of the Project").

Testing of the modified framework was accomplished by both programmatic means and through person-in-the-loop trials using data sources as outlined in the original proposal. Of the nearly 15,000 papyri cataloged (representing nearly all published Greek and Latin literary papyri), metadata for 14,613 was imported from the Leuven Database of Ancient Books. Some metadata was subsequently enhanced and refined, both programmatically and through the Papyrological Editor, drawing on additional sources and on printed editions. Additional metadata records were augmented with information from the Spanish DVCTVS collection, a collaboration anticipated in the proposal.

Edition texts were also added by import and through-the-web. Team members encoded over 900 transcriptions of Greek papyri, totalling more than 110,000 lines of text. This total represents content provided by Heidelberg staff, volunteers, and the the project teams of both the ERC-funded "DIGMEDTEXT - Online Humanities Scholarship: A Digital Medical Library based on Ancient Texts" at the University of Parma in Italy and the Anagnosis project at the University of Würzburg. This number is very close to the anticipated figure of 1,100; moreover, several of the editions (particularly those from Herculaneum contributed by Anagnosis) were especially lengthy and complex, providing a variety of challenges. The metadata and editions are available to the public not only through the litpap.info interface, but also as raw data published under open license on the Github platform.

## Audiences

Because most literary papyri have been published only in specialist literature and are known to a limited circle, DCLP takes the first steps toward considerably enlarging their potential audience. DCLP's primary audience is naturally those scholars who work with Greek and Latin literatures (e.g., classical philologists, theologians, ancient historians), but the technology will enable extension to other ancient literatures in subsequent phases as papyri.info has demonstrated for the documentary papyri. The eventual inclusion of translations, especially for texts not part of the standard authors transmitted by medieval manuscripts (and generally available in translation) will make materials in the DCLP more widely usable by researchers in other fields, whether historians (of law, religion, medicine, science, philosophy, etc.) or scholars in other fields of literature. This accessibility will be particularly important in areas like ancient medicine, astronomy, and mathematics, where many of the papyri have never been readily available in translation.

## Evaluation

By design, evaluation was integral to project activity through the period of performance. It was focused on identifying needs and difficulties of early adopters and event participants and translating these into actionable feature requests that could be articulated in the tracker, prioritized, and implemented by the development teams.

## Continuation of the Project

The end of the funding period does not constitute an end to the project. We had always imagined this as a first step in a community-driven initiative that would steer ongoing content creation and oversee quality control. In an effort to increase the number both of transcriptions of literary texts and of system users, we plan to offer more webinars on subjects that intersect with work on the corpus. Moreover, we have entered informal partnerships with papyrologists who have an interest in seeing the textual sources that they work on made available in DCLP. For example, the Heidelberg team is working with Maria Serena Funghi and Valeria Piano, who are contributing both metadata and texts of papyri published in the *Corpus dei papiri filosofici greci e latini* series. Dr. Piano used some of this material for the DCLP training session in New York in August, which she co-instructed.

In addition to content creation, there are system-related issues that will need to be addressed, in some cases in the short term. Two issues in particular could not be resolved in the first round of funding. The first is full implementation of the translation tool. At the moment, modern-language translations can be entered but not processed by the editor. This means that, although they reside in the XML, they are not being surfaced in the HTML. Once the translation editor is fully operational, however, we can begin to expand the user base by offering English, German and other translations of the ancient texts. The second system-related feature that needs attention is the tool that allows creation of new metadata (i.e., metadata that postdates the data transfer from LDAB). We did not appreciate at the outset the fact that we would need to build into the editor the necessary mechanism for creating records of newly published literary texts. As discussed above, cooperation with Leuven gave us an initial set of metadata files for ca. 15,000 objects but we were not able to automate the process of delivering newly created LDAB content to Heidelberg in a repeatable manner. Thus, any information that has been added to the LDAB since the transfer of data from Leuven to Heidelberg a couple of years ago cannot be viewed in DCLP. Further, we did not have time to implement a tool for creating new metadata from scratch through the online interface. Addressing these needs is of urgency and Heidelberg is working to secure funding in the near future in order to do so.

## Long-term Impact

Papyrus was the primary writing material in antiquity and perhaps Egypt's most important legacy. The papyri are the most significant written records we have from the ancient Mediterranean world that cover the entire range of private and public life. On these materials were recorded everything from high literature to the documents and other communications of daily life in a variety of scripts and languages reflecting the mosaic of cultures of ancient Egypt,



with Greek, Egyptian, and, later, Arabic being the predominant languages. Texts in Latin, Aramaic, and Pahlavi are less common but still significant.

Roughly ten percent of texts studied to date (numbering altogether about 70,000 in Greek and Latin by recent estimate) are fragments of literature, constituting either far more ancient witnesses to works known otherwise from medieval manuscripts or "new" texts not preserved elsewhere. Literary papyri reveal the state of literary texts before the Middle Ages, which through the process of copying both transmitted and contaminated the ancient textual tradition. These papyri are also the source of most of the additions to our knowledge of ancient literature in the past 125 years and the basis for many important new ideas about ancient literacy, reading culture, and textual transmission. Important works preserved exclusively in papyri include the lyrics of Sappho and Archilochus and the Paeans of Pindar, the verses of Callimachus, the comedies of Menander, the Constitution of the Athenians by Aristotle, lost plays of the classical Athenian dramatists, and early Christian and Gnostic works which once competed with the New Testament, not to mention a host of unattributed works in a great variety of genres. From Herculaneum, one of the towns buried in AD 79 by Mount Vesuvius, comes an entire ancient philosophical library on carbonized papyrus. On the most generous definition, however, "literary" papyri include an even richer harvest, such as recipes for medicines, horoscopes, geometry problems, magical spells, tables of fractions, conjugations of verbs, and exercises from ancient schools.

In the past, literary papyrology has often served mainly as an auxiliary means for verifying and supplementing literature preserved in MS traditions. Texts whose authorship has not been established have sometimes been ignored, making it more difficult to identify and classify previously unknown literary works. In short, literary papyri have often not been appreciated for what they actually were: ancient books written by and distributed among members of Greco-Roman society and cornerstones of ancient learning, entertainment, medicine, religion, etc. Viewing the whole literary corpus not only as parts of a stemma in a textual tradition produced by the application of traditional philological methods, but also as cultural artifacts created in a specific historical setting enriches our understanding of ancient society considerably. Moreover, literary papyrology has the potential further to transform our knowledge of the development of the Greek and Latin languages in everyday use, which bears not only on historical linguistics, but on interpretation of Jewish and Christian sacred texts.

A comprehensive, searchable collection of literary texts as texts will put the corpus of literary and subliterate papyri at the very fingertips of researchers around the world, significantly reducing the time needed to track down relevant textual evidence. It will offer scholars and students in places that do not have library resources on the scale of large research centers better opportunities to work on this material. The impact of free online availability of the documentary papyri on the revival of work in papyrology in Egypt, for example, has been notable, and we expect a similar result as DCLP matures. The material will also be important for a broader public, since the inclusion of translations along with the ancient texts in the original languages will provide non-specialists with unprecedented access. Finally, since we are encoding the literary texts and metadata with EpiDoc (i.e., with the same methods and data

formats used not only for the documentary papyri, but for a wide variety of ancient texts from diverse cultures and disciplines), they will be searchable alongside other corpora, facilitating work in the areas of social history, linguistics, ancient education, medicine, and related areas.

## Grant Products and Outreach

Project code and data are openly available from <https://github.com/DCLP>. The development demonstration version of the DCLP application is openly available at <http://litpap.info>.